### **1. Source Data Description:**

The source data captures breast cancer patient information, their visits, procedures, providers, conditions, care sites, and medical device distributors. Key information includes patient demographics, visit details, performed procedures, attending providers, medical conditions, care site information, and details about medical device distributors for certain procedures.  
  
Source data link (<https://app.snowflake.com/marketplace/listing/GZ1M6ZXHO15/precision-data-datasus-access-real-world-data-modeling-free-sample?search=datasus>)

### **2. Why Data Vault in Healthcare:**

* **Scalability and Flexibility:**
  + Adaptable architecture supports evolving healthcare data sources and analytic requirements without major restructuring.
* **Traceability and Auditing:**
  + Design principles ensure clear traceability of changes over time, vital for auditing and compliance.
* **Handling Slowly Changing Dimensions (SCDs):**
  + Manages changes in dimension attributes over time, aligning with the dynamic nature of healthcare data.
* **Linking Disparate Data Sources:**
  + Integrates diverse healthcare data sources through a systematic and traceable approach to data linking.
* **Agility in Data Integration:**
  + Agile framework accommodates dynamic healthcare environments, facilitating quick adaptation to changes and new data sources.

### **3. Data Vault Model**

#### Hubs:

* **HUB\_PATIENT:**
  + **Attributes:** PERSON\_BK (Business Key), PERSON\_SK (Surrogate Key), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **HUB\_VISIT:**
  + **Attributes:** VISIT\_OCCURRENCE\_BK (Business Key), VISIT\_SK (Surrogate Key), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **HUB\_PROCEDURE:**
  + **Attributes:** PROCEDURE\_BK (Business Key), PROCEDURE\_SK (Surrogate Key), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **HUB\_PROVIDER:**
  + **Attributes:** PROVIDER\_BK (Business Key), PROVIDER\_SK (Surrogate Key), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **HUB\_CONDITION:**
  + **Attributes:** CONDITION\_BK (Business Key), CONDITION\_SK (Surrogate Key), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **HUB\_CARE\_SITE:**
  + **Attributes:** CARE\_SITE\_BK (Business Key), CARE\_SITE\_SK (Surrogate Key), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **HUB\_MED\_DEVICE\_DISTRIBUTOR:**
  + **Attributes:** MED\_DEVICE\_DISTRIBUTOR\_BK (Business Key), MED\_DEVICE\_DISTRIBUTOR\_SK (Surrogate Key), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).

#### Links:

* **LINK\_PATIENT\_VISIT:**
  + **Attributes:** LINK\_SK (Surrogate Key), PERSON\_SK (Surrogate Key), VISIT\_SK (Surrogate Key), START\_DATE (Start Date), END\_DATE (End Date), IS\_CURRENT (Is Current), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **LINK\_VISIT\_PROCEDURE:**
  + **Attributes:** LINK\_SK (Surrogate Key), VISIT\_SK (Surrogate Key), PROCEDURE\_SK (Surrogate Key), START\_DATE (Start Date), END\_DATE (End Date), IS\_CURRENT (Is Current), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **LINK\_VISIT\_PROVIDER:**
  + **Attributes:** LINK\_SK (Surrogate Key), VISIT\_SK (Surrogate Key), PROVIDER\_SK (Surrogate Key), START\_DATE (Start Date), END\_DATE (End Date), IS\_CURRENT (Is Current), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **LINK\_PATIENT\_CONDITION:**
  + **Attributes:** LINK\_SK (Surrogate Key), PERSON\_SK (Surrogate Key), CONDITION\_SK (Surrogate Key), START\_DATE (Start Date), END\_DATE (End Date), IS\_CURRENT (Is Current), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).
* **LINK\_VISIT\_CARE\_SITE:**
  + **Attributes:** LINK\_SK (Surrogate Key), VISIT\_SK (Surrogate Key), CARE\_SITE\_SK (Surrogate Key), START\_DATE (Start Date), END\_DATE (End Date), IS\_CURRENT (Is Current), LOAD\_TIMESTAMP (Load Timestamp), RECORD\_SOURCE (Record Source).

#### Satellites:

* **SAT\_PATIENT:**
  + **Attributes:** PERSON\_SK (Surrogate Key), AGE, GENDER, PATIENT\_CITY, LOAD\_TIMESTAMP, RECORD\_SOURCE, START\_DATE, END\_DATE, IS\_CURRENT, (Other descriptive attributes related to patients).
* **SAT\_VISIT:**
  + **Attributes:** VISIT\_SK (Surrogate Key), VISIT\_TYPE, VISIT\_START\_DATE, VISIT\_END\_DATE, INPATIENT\_BED\_TYPE, ICU\_DAYS, INPATIENT\_DAYS, LOAD\_TIMESTAMP, RECORD\_SOURCE, START\_DATE, END\_DATE, IS\_CURRENT, (Other descriptive attributes related to visits).
* **SAT\_PROCEDURE:**
  + **Attributes:** PROCEDURE\_SK (Surrogate Key), PROCEDURE\_NAME, PROCEDURE\_CLASS\_NAME, PROCEDURE\_DOMAIN\_NAME, PROC\_QT, PROC\_PAID, PROC\_YEAR, PROC\_MONTH, LOAD\_TIMESTAMP, RECORD\_SOURCE, START\_DATE, END\_DATE, IS\_CURRENT, (Other descriptive attributes related to procedures).
* **SAT\_PROVIDER:**
  + **Attributes:** PROVIDER\_SK (Surrogate Key), PROVIDER\_NAME, CRM, PROVIDER\_CITY, PROVIDER\_STATE, LOAD\_TIMESTAMP, RECORD\_SOURCE, START\_DATE, END\_DATE, IS\_CURRENT, (Other descriptive attributes related to providers).
* **SAT\_CONDITION:**
  + **Attributes:** CONDITION\_SK (Surrogate Key), CONDITION\_NAME, CONDITION\_CLASS\_NAME, CONDITION\_DOMAIN\_NAME, LOAD\_TIMESTAMP, RECORD\_SOURCE, START\_DATE, END\_DATE, IS\_CURRENT, (Other descriptive attributes related to conditions).
* **SAT\_CARE\_SITE:**
  + **Attributes:** CARE\_SITE\_SK (Surrogate Key), CARE\_SITE\_NAME\_ADM, CARE\_SITE\_NAME, CARE\_SITE\_CITY, CARE\_SITE\_STATE, CARE\_SITE\_REGION, LOAD\_TIMESTAMP, RECORD\_SOURCE, START\_DATE, END\_DATE, IS\_CURRENT, (Other descriptive attributes related to care sites).
* **SAT\_MED\_DEVICE\_DISTRIBUTOR:**
  + **Attributes:** MED\_DEVICE\_DISTRIBUTOR\_SK (Surrogate Key), MED\_DEVICE\_DISTRIBUTOR\_NAME\_ADM, MED\_DEVICE\_DISTRIBUTOR\_NAME, MED\_DEVICE\_DISTRIBUTOR\_LEGAL\_TYPE, MED\_DISTRIBUTOR\_ECONOMY\_TYPE, MED\_DISTRIBUTOR\_CITY, MED\_DISTRIBUTOR\_STATE, LOAD\_TIMESTAMP, RECORD\_SOURCE, START\_DATE, END\_DATE, IS\_CURRENT, (Other descriptive attributes related to medical device distributors).

### **3. Data Analytics:**

#### Business Use Cases:

#### **Patient Demographics Analysis:**

* Understand the demographics of patients, including age distribution, gender distribution, and geographic locations. This can help in targeted healthcare services and resource allocation.

**Patient History Analysis:**

* Identify the historical details of a patient, including past visits, procedures, and conditions.

**Most Frequent Procedures:**

* Identify the most performed procedures. This information can guide resource planning, equipment acquisition, and training programs for healthcare professionals.

**Patient Journey Mapping:**

* Map the patient's journey from initial visit to follow-up care. This can help in identifying bottlenecks, improving patient experience, and optimizing care pathways.